

# Supervised Syntax-based Alignment between English Sentences and Abstract Meaning Representation Graphs

Chenhui Chu<sup>1</sup> and Sadao Kurohashi<sup>2</sup>

<sup>1</sup>Japan Science and Technology Agency

<sup>2</sup>Graduate School of Informatics, Kyoto University

chu@pa.jst.jp, kuro@i.kyoto-u.ac.jp

## Abstract

As alignment links are not given between English sentences and Abstract Meaning Representation (AMR) graphs in the AMR annotation, automatic alignment becomes indispensable for training an AMR parser. Previous studies formalize it as a string-to-string problem and solve it in an unsupervised way, which suffers from data sparseness due to the small size of training data for English-AMR alignment. In this paper, we formalize it as a syntax-based alignment problem and solve it in a supervised manner based on syntax trees, which can address the data sparseness problem by generalizing English-AMR tokens to syntax tags. Experiments verify the effectiveness of the proposed method not only for English-AMR alignment, but also for AMR parsing.

## 1 Introduction

Abstract Meaning Representation (AMR) is a sentence level semantic annotation, which is represented in a rooted, directed, and edge-labeled graph [Banarescu *et al.*, 2013]. Nodes of a graph are *concepts* (e.g., “possible” in Figure 1), while edges are labeled with semantic *roles* (e.g., “:ARG4” in Figure 1). AMR concepts consist of predicate senses, named entities, and lemmas of English tokens. AMR roles consist of core semantic roles from the Propbank [Palmer *et al.*, 2005] and fine-grained semantic roles defined specifically for AMR. As AMR annotation has no explicit alignment with the tokens in the English sentence, automatic alignment becomes a requirement for training AMR parsers [Flanigan *et al.*, 2014; Wang *et al.*, 2015; Werling *et al.*, 2015; Artzi *et al.*, 2015; Pust *et al.*, 2015; Zhou *et al.*, 2016; Misra and Artzi, 2016; Peng *et al.*, 2017].

The alignment problem between English sentences and AMR graphs is not trivial. There are two reasons. Firstly, the problem itself is complicated. Concepts do not always have a direct matching among the English tokens in a given sentence. For example, in Figure 1 the English token “could” is represented as the concept “possible,” and aligning them is not easy. It becomes more difficult in the English token-to-role alignment case. For example, in Figure 1, we should align the English token “to” to the role “:ARG4.” Secondly,

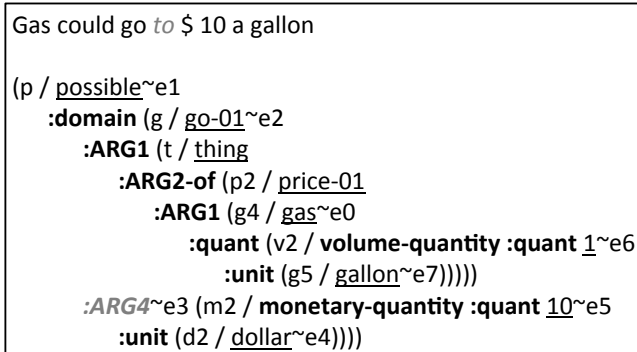


Figure 1: An example of English-AMR alignment (AMR concepts are underlined, AMR roles are in bold, ~e denotes alignment and the numbers after ~e are English token indices).

the training data for English-AMR alignment is very small. The biggest publicly available data set only contains 13,050 English-AMR pairs, which is significantly smaller than conventional alignment settings in machine translation (MT).

[Flanigan *et al.*, 2014] is the first study of English-AMR alignment, which proposes a rule-based method. The limitation of the rule-based method is that it cannot benefit from more annotation of English-AMR pairs. A data driven alignment method also has been proposed [Pourdamghani *et al.*, 2014]. They formalize the string-to-graph alignment problem as a string-to-string problem by linearizing the AMR graph. Then they apply the conventional unsupervised string-to-string alignment models (i.e., IBM models [Brown *et al.*, 1993]) for this problem. However, this method significantly suffers from data sparseness due to the small size of training data.

Using syntax trees in a supervised manner has shown its effectiveness in the alignment problem in MT [Riesa *et al.*, 2011]. Motivated by this, in this paper, we formalize the English-AMR graph alignment problem as a syntax-based alignment problem. We then apply the supervised syntax-based alignment model [Riesa *et al.*, 2011] for this. Our proposed method generalizes pure English-AMR tokens to syntax tags, which can address the data sparseness problem of the previous study [Pourdamghani *et al.*, 2014].

Experiments conducted on the benchmark dataset show that our proposed method outperforms the unsupervised alignment model of [Pourdamghani *et al.*, 2014] by 1.7% absolute F-score on alignment accuracy, although it is trained on only 100 English-AMR pairs that are annotated with gold alignments. Using the alignments by our proposed method instead of the unsupervised alignments for a state-of-the-art AMR parser [Pust *et al.*, 2015] improves the parsing accuracy of 0.4% absolute Smatch F-score [Cai and Knight, 2013].

## 2 Related Work

Three different alignment criteria, and their corresponding alignment methods have been proposed [Flanigan *et al.*, 2014; Pourdamghani *et al.*, 2014; Werling *et al.*, 2015]. [Flanigan *et al.*, 2014] is the most prior work of English-AMR alignment. They proposed an alignment criterion that aligns a span of English tokens to an AMR graph concept fragment. This means that in Figure 1, e.g., the English token “gas” should be aligned to the graph fragment “(t / thing :ARG2-of (p2 / price-01 :AMR1 (g4 / gas.” Their criterion, however, does not align AMR roles explicitly. [Flanigan *et al.*, 2014] proposed a rule-based method for this type of alignment. In contrast, [Pourdamghani *et al.*, 2014] proposed an alignment criterion that not only aligns AMR concepts but also roles to English tokens explicitly. In addition, in their criterion, each concept and role is aligned to at most one English token, but each English token can be aligned to many concepts/roles. An example of the alignment criterion of [Pourdamghani *et al.*, 2014] is shown in Figure 1. They proposed an unsupervised string-to-string alignment model for this. [Werling *et al.*, 2015] essentially adopted the alignment criterion of [Pourdamghani *et al.*, 2014], except that they forced every AMR concept aligns to some English tokens. They proposed a boolean linear programming method for this. Note that the accuracies of these studies were reported on different golden alignment data, and thus they are not directly comparable. We adopt the alignment criterion of [Pourdamghani *et al.*, 2014], and directly compare the alignment accuracy with their method.

For AMR parsing, [Flanigan *et al.*, 2014] is also the most prior work. They proposed a graph based parsing method that finds a maximum spanning and connected subgraph via structured prediction for AMR parsing. Their parser is publicly available as the JAMR parser.<sup>1</sup> [Werling *et al.*, 2015] extended the study of [Flanigan *et al.*, 2014] by proposing generative actions for subgraph derivation based on their alignment criterion. [Zhou *et al.*, 2016] extended the study of [Flanigan *et al.*, 2014] by proposing a beam search algorithm. [Wang *et al.*, 2015] proposed a transition based method that first parses English sentences to dependency trees and then transforms the dependency trees to AMR graphs. Their parser is publicly available as the CAMR parser.<sup>2</sup> [Artzi *et al.*, 2015] proposed using the combinatory categorial grammar (CCG) for

AMR parsing. [Misra and Artzi, 2016] developed the CCG AMR parsing method of [Artzi *et al.*, 2015] based on neural networks. Note that [Flanigan *et al.*, 2014; Wang *et al.*, 2015; Artzi *et al.*, 2015; Zhou *et al.*, 2016; Misra and Artzi, 2016] adopted the alignment criterion and method of [Flanigan *et al.*, 2014]. [Pust *et al.*, 2015] treated AMR parsing as a string-to-tree, syntax-based MT problem. After transforming English sentences to trees, they further convert the trees to AMR graphs. Their parser is publicly available as the ISI AMR parser.<sup>3</sup> Sequence-to-sequence based AMR parsing also has been proposed [Peng *et al.*, 2017], however it suffers from data sparseness due to the small size of AMR training data. Both [Pust *et al.*, 2015] and [Peng *et al.*, 2017] adopted the alignment criterion and method of [Pourdamghani *et al.*, 2014]. Note that the parsers of [Werling *et al.*, 2015; Zhou *et al.*, 2016; Peng *et al.*, 2017] are not publicly available. We apply the alignments by our proposed method to the ISI AMR parser [Pust *et al.*, 2015], and also compare the parsing performance with the other publicly available parsers (i.e., JAMR and CAMR).

## 3 Baseline Alignment Method

The baseline method that we compare to is the ISI alignment [Pourdamghani *et al.*, 2014]. The ISI alignment method formalizes the English-AMR graph alignment problem as a string-to-string alignment problem, by linearizing an AMR graph to a string. It includes three steps: preprocessing, string-to-string alignment, and postprocessing.

### 3.1 Preprocessing

- Linearize the AMR using a depth-first traversal. For example, the AMR graph in Figure 1 will be linearized to “possible :domain go-01 :ARG1 thing :ARG2-of price-01 :ARG1 gas :quant volume-quantity :quant 1 :unit gallon :ARG4 monetary-quantity :quant 10 :unit dollar.”
- Remove the tokens that are rarely aligned, to improve the precision with a small sacrifice of recall. On the English side, this removes stop words, such as articles “a”, “an”, “the”; On the AMR side, this removes special concepts, and roles, such as “:arg0”, “:quant”, “:op1” that do not usually align, quotes, and sense tags. After this step, the English sentence in Figure 1 becomes “Gas could go to \$ 10 gallon”; the AMR is transferred to “possible :domain go thing :arg2-of price gas 1 gallon :arg4 10 dollar”.
- Lowercase and stem both sides to the first four letters. This is necessary to address the sparseness of the training data, which is very small compared to the size of the training data for conventional word alignment of MT. This converts English to “gas coul go to \$ 10 gall”, and AMR to “poss :domain go thin :arg2-of pric gas 1 gall :arg4 10 doll” in Figure 1.

### 3.2 String-to-String Alignment

As the preprocessing step has converted the English-AMR graph alignment problem to a string-to-string align-

<sup>1</sup><http://github.com/jflanigan/jamr>

<sup>2</sup><https://github.com/c-amr/camr>

<sup>3</sup><http://www.isi.edu/~pust/amrparser.tar.gz>

ment problem, the widely used IBM alignment models [Brown *et al.*, 1993] that are based on token sequences can be applied. To further improve the alignment accuracy, [Pourdamghani *et al.*, 2014] also proposed a symmetrization constraint that encourages agreement of the parameter learning in two directions for the IBM models.

### 3.3 Postprocessing

The string-to-string alignments are finally projected back to the original English sentence and the AMR graph to obtain English-AMR graph alignments. This can be done easily by memorizing the corresponding token positions before and after the preprocessing.

## 4 Proposed Alignment Method

We use the same pipeline as [Pourdamghani *et al.*, 2014], however, we formalize it as a constituency tree based alignment problem, and apply the hierarchical alignment model of [Riesa *et al.*, 2011]. This method has been proposed for conventional word alignment of MT, however, it has not been used for English-AMR graph alignment.

## 4.1 Constituency Trees for English and AMR

Constituency trees for English can be obtained via a conventional syntactic parser. In this study, we parse original English sentences with the Berkeley parser<sup>4</sup> [Petrov and Klein, 2007]. We process obtained constituency trees by discarding the stop words, and replacing the leaf tokens with their stems. An example of the final tree is shown in Figure 2.

For AMR, we convert AMRs to constituency trees using the method proposed in [Pust *et al.*, 2015] with the following steps:

- Arbitrarily disconnect multiple parents from each node.
- Propagate the edge labels (roles) to leaves, and add pre-terminals X.
- Restructure the tree with role labels as intermediates.

We do not apply the reordering steps, because they requires alignments. For more details of these steps, please refer to [Pust *et al.*, 2015]. We used the AMR to syntax tree conversion code provided by [Pust *et al.*, 2015] for the above conversion.

We then process the converted AMR trees by discarding special concepts and roles that are rarely aligned, and replacing leaf tokens with their stems. Note that the converted AMR trees usually are not isomorphic to the English trees. For example, “could” is the grandchild of the root in the English tree, while in the converted AMR tree “possible” is the direct child of the root.

## 4.2 Hierarchical Alignment on Constituency Trees

Figure 2 shows an overview of the application of the hierarchical alignment model [Riesa *et al.*, 2011] to our problem. The model hierarchically searches for the k-best alignment by constructing partial alignments over a target constituency tree,<sup>5</sup> in a bottom-up manner (from leaf nodes to the root).

<sup>4</sup><https://github.com/slavpetrov/berkeleyparser>

<sup>5</sup>The source side could be either a constituency tree or a token sequence.

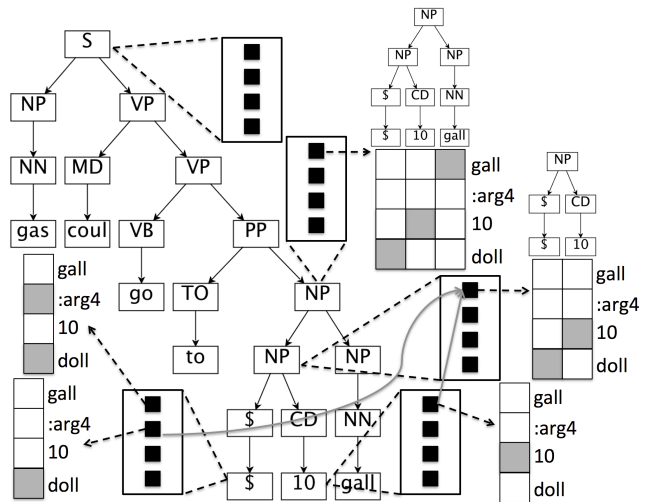


Figure 2: Proposed alignment method (Each black square represents a partial alignment; each gray square represents an alignment link in an alignment matrix).

Each node in the tree has partial alignments, which are sorted by alignment scores. A partial alignment for a node is an alignment matrix of AMR tokens or *null*, covered by the node, and it is represented as a black square. We only keep a beam size of  $k$  for partial alignments for each node,<sup>6</sup> to reduce the computational cost. For example, in Figure 2 the beam size  $k = 4$ . Firstly, 4-best partial alignments are generated for all the leaf nodes. These partial alignments are then linearly combined to generate partial alignments for the non-terminals in the constituency tree. For example, the partial alignments of the leaf node “\$” and “10” are combined to generate 4-best partial alignments for the node “NP”. We hierarchically perform this process until we reach the root node.

One important merit of this model is that it is a discriminative model that can incorporate various features including local and non-local features. Local features are the ones that can be factored among the local productions in a tree, and otherwise they are non-local features. Local features include source syntactic, target syntactic, source-target joint syntactic, translation rule and same token features. Non-local features include lexical translation probabilities, and third party alignment features.

The score of a partial alignment is a linear combination of these features by their weights. The weights of the features are learnt against a set of pairs with gold alignments, using the online averaged perceptron algorithm [Collins, 2002]. The learning objective is defined as:

$$\hat{y} = \arg \max_{y \in Y(x_i)} L(y_i, y) + w \cdot h(y) \quad (1)$$

where  $x_i$  is a token sequence pair and their parse trees;  $y_i$  is the gold alignment for  $x_i$ ;  $Y(x_i)$  denotes all the possible alignment outputs for  $x_i$ ;  $w$  is a weight vector;  $h(y)$  is a vector of feature values;  $L(y_i, y)$  is a loss function to measure how bad it would be to guess  $y$  instead of  $y_i$ , which is defined

<sup>6</sup>We used a beam size of 128 in our experiments.

Original split			
	train	dev	test
# pairs	10,311	1,368	1,371
# AMR tokens	364k	48.9k	51.0k
# AMR roles	177k	23.7k	24.8k
# English tokens	213k	28.8k	29.5k

Our split			
	train	dev	test
# pairs	10,311+200	1,368-100	1,371-100
# AMR tokens	370k	45.1k	48.6k
# AMR roles	180k	21.9k	23.6k
# English tokens	217k	26.5k	27.8k

Table 1: Statistics of the AMR corpus for AMR parsing.

	train	dev	test
# pairs	100	50	50
# AMR tokens	3.0k (54.7%)	1.5k (51.6%)	1.6k (52.2%)
# AMR roles	1.5k (23.5%)	0.7k (21.0%)	0.8k (21.6%)
# English tokens	2.0k (74.9%)	0.9k (75.8%)	1.1k (75.4%)

Table 2: Statistics of the gold alignment data (The numbers in parentheses are the percentages of the tokens aligned in the gold alignment data).

as:

$$L(y_i, y) = 1 - F_1(y_i, y) \quad (2)$$

where  $F_1(y_i, y)$  is a F-score.  $w$  is updated at each iteration as:

$$w = w + h(y_i) - h(\hat{y}) \quad (3)$$

We stop the update of  $w$  when a predefined iteration number has been reached. After the learning of a set of  $w$  at each iteration, we select the  $w$  that achieves the best F-score on a development set for decoding. For more details of the features and the learning algorithm, please refer to [Riesa *et al.*, 2011].

## 5 Experiments

We conducted both alignment and AMR parsing experiments, to verify the effectiveness of our proposed alignment method.

### 5.1 Settings

The data used in our experiments was the Linguistic Data Consortium AMR corpus release 1.0 (LDC2014T12), consisting of 13,050 AMR/English sentence pairs.<sup>7</sup> The statistics of the original split in the AMR corpus for training, development, and testing AMR parsers are shown in upper part of Table 1. Among which, 100 development<sup>8</sup> and 100 testing<sup>9</sup> pairs were manually annotated with gold alignments [Pourdamghani *et al.*, 2014]. As these 200 pairs were used for training and testing (and tuning for our proposed alignment model) alignment models in our study, we moved them

<sup>7</sup><https://catalog.ldc.upenn.edu/LDC2014T12>

<sup>8</sup><http://www.isi.edu/natural-language/mt/dev-gold.txt>

<sup>9</sup><http://www.isi.edu/natural-language/mt/test-gold.txt>

Type	Method	Precision	Recall	F-score
Concept	ISI	96.2%	85.6%	90.6%
	Proposed	<b>97.1%</b>	<b>87.8%</b>	<b>92.2%</b> †
	Upper bound	99.7%	94.4%	97.0%
Role	ISI	69.1%	43.9%	53.7%
	Proposed	<b>69.7%</b>	<b>47.6%</b>	<b>56.6%</b> †
	Upper bound	95.4%	66.1%	78.1%
Concept + Role	ISI	92.0%	77.1%	83.9%
	Proposed	<b>92.7%</b>	<b>79.6%</b>	<b>85.6%</b> †
	Upper bound	99.0%	88.6%	93.5%

Table 3: Alignment results (“Proposed” shows the best results among the different syntax usages, “Upper bound is” the upper bound after removing stop words in English and rarely aligned concepts, roles in AMR, “†” indicates that the F-score is significantly better than “ISI” at  $p < 0.01$ ).

to the training data for our AMR parsing experiments. The statistics of our split of the AMR corpus for AMR parsing is shown in lower part of Table 1. We further mixed these 200 pairs with gold alignments pair by pair, and split them into 100, 50, 50 for training, tuning, and testing, respectively in our alignment experiments. Table 2 shows the statistics of the gold alignment data.

For alignment, we compared our proposed alignment method with the baseline method of [Pourdamghani *et al.*, 2014]. For the baseline method, we ran the publicly available toolkit ISI aligner<sup>10</sup> on the training data of our split in Table 1. The ISI aligner is an implementation of the method described in [Pourdamghani *et al.*, 2014]. For our proposed method, we trained and tuned the alignment model on the 100 training and 50 development pairs, respectively, with the open source supervised alignment toolkit Nile<sup>11</sup> [Riesa *et al.*, 2011]. As the third party alignment feature for Nile, we used the trained ISI alignments. Lexical translation probabilities were generated from the ISI alignments on the training data of our split in Table 1. Alignment results were reported on the 50 testing pairs. In addition, we compared different ways of using syntax trees for our proposed method:

- AMR(string)-En(tree): Use AMR strings as the source side, and English trees as the target side.
- AMR(tree)-En(tree): Use converted AMR trees as the source side, and English trees as the target side.
- En(tree)-AMR(tree): Use English trees as the source side, and converted AMR trees as the target side.
- Grow-diag-final-and (1): A symmetrization of the alignment results of AMR(string)-En(tree) and En(tree)-AMR(tree) with the grow-diag-final-and heuristic [Och and Ney, 2003], which is commonly used in MT.
- Grow-diag-final-and (2): A symmetrization of the alignment results of AMR(tree)-En(tree) and En(tree)-AMR(tree) with the grow-diag-final-and heuristic [Och and Ney, 2003].

For AMR parsing, we compared the parsing performance of the state-of-the-art AMR parser [Pust *et al.*, 2015] using

<sup>10</sup><http://www.isi.edu/damghani/papers/Aligner.zip>

<sup>11</sup><https://github.com/neubig/nile>



Type	Method	Precision	Recall	F-score
Concept	AMR(string)-En(tree)	96.1%	87.5%	91.6%
	AMR(tree)-En(tree)	95.9%	87.2%	91.4%
	En(tree)-AMR(tree)	94.7%	88.0%	91.3%
	Grow-diag-final-and (1)	94.7%	<b>88.6%</b>	91.6%
	Grow-diag-final-and (2)	<b>97.1%</b>	87.8%	<b>92.2%</b>
Role	AMR(string)-En(tree)	<b>78.8%</b>	43.3%	55.9%
	AMR(tree)-En(tree)	77.3%	43.3%	55.5%
	En(tree)-AMR(tree)	61.7%	48.6%	54.4%
	Grow-diag-final-and (1)	67.8%	<b>50.2%</b>	<b>57.7%</b>
	Grow-diag-final-and (2)	69.7%	47.6%	56.6%
Concept+Role	AMR(string)-En(tree)	<b>93.8%</b>	78.5%	85.5%
	AMR(tree)-En(tree)	93.4%	78.3%	85.2%
	En(tree)-AMR(tree)	88.8%	80.0%	84.2%
	Grow-diag-final-and (1)	90.2%	<b>80.8%</b>	85.2%
	Grow-diag-final-and (2)	92.7%	79.6%	<b>85.6%</b>

Table 4: Syntax usage comparison results for our proposed method.

either the baseline ISI alignments or our proposed alignments, respectively. We trained the baseline alignment model using the same method described above. Alignments for the parsing training data are available once we trained the baseline alignment model. For our proposed alignment method, we trained the alignment model with the best syntax usage and further applied the trained model on the parsing training data for obtaining alignments. In addition, we compared with the performance of other public available AMR parsers [Flanigan *et al.*, 2014; Wang *et al.*, 2015] that use alignments with different annotation criteria from ours, by running their parsers with default settings. All AMR parsing experiments were conducted on our split of the AMR corpus in Table 1.

## 5.2 Alignment Results

Table 3 shows the alignment results. We report the alignment accuracies for the concept, role and both types of tokens, respectively. The significance tests were performed using the bootstrapping method [Zhang *et al.*, 2004]. We can see that our proposed method outperforms the ISI alignment for all the alignment types. However, there is still a gap between it and the upper bound, especially for role alignment. Although they are not directly comparable, the concept precision 97.1% and F-score 92.2% obtained by our proposed method are also significantly higher than the concept precision 83.2% reported in [Werling *et al.*, 2015] and the concept F-score 90% reported in [Flanigan *et al.*, 2014], respectively.

Table 4 shows the results of using syntax trees in different ways for the proposed method. AMR(tree)-English(tree) performs slightly worse than AMR(string)-English(tree), due to the bad isomorphism between converted AMR trees and English trees. Using converted AMR trees as the target side seems to be a bad idea for the precision, but it is helpful for improving the recall. The reason for this is that English-to-AMR is a one-to-many alignment problem, while En(tree)-AMR(tree) could produce many-to-one alignments for English-to-AMR due to the peculiarities of Nile, which decreases the precision but improves the recall. The grow-diag-final-and heuristic leverages both the high precision of AMR(tree)-En(tree) and the high recall of En(tree)-AMR(tree), and thus further improves the F-score; it also

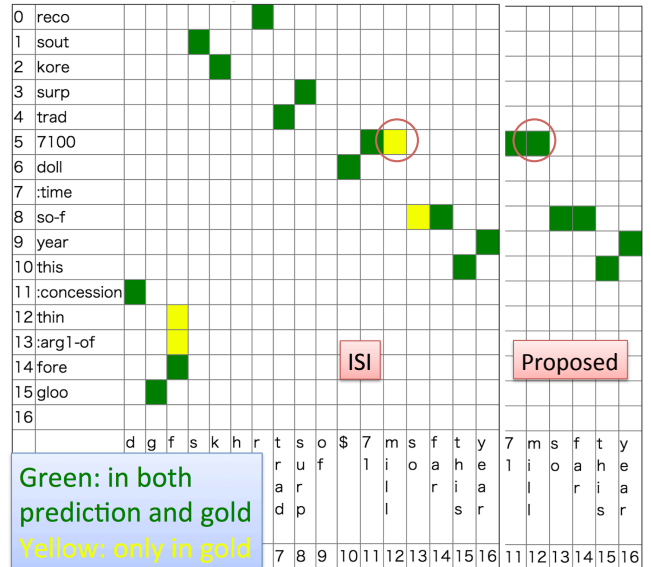


Figure 3: An alignment comparison example of better concept recall.

improves the role F-score by using the grow-diag-final-and heuristic for AMR(string)-En(tree) and En(tree)-AMR(tree).

To further understand the reason for the alignment improvement, we investigated many alignment examples. We found that the main reason for the alignment improvement is the generalization of pure English-AMR tokens to syntax tags, which addresses the data sparseness problem. Figure 3 shows an alignment example of better concept recall by our proposed method. Comparing the ISI alignment and our proposed method, we can see that our proposed method successfully aligns the AMR concept “7100(000)” to the English tokens “71” and “mill(ion),” while the ISI alignment fails. “7100(000)” is not aligned to “mill(ion),” due to the sparseness of the training data. With the help of syntax information, both “7100(000)” and “mill(ion)” are generalized to cardinal numbers “Aquant” and “QP (CD),” respectively. As cardinal numbers co-occur many times in the training data, our proposed method learns a big positive weight, which successfully aligns them.

Figure 4 shows an alignment example of better role recall by our proposed method. We can see that our proposed method successfully aligns the AMR role token “:topic” to the English token “with,” while the ISI alignment fails. This is again improved by the syntax information. As “:topic” and “with” do not co-occur in the training data, it is very difficult for the ISI alignment model to align them. Our proposed method generalizes them to “Ctopic” and “IN,” respectively. For which, a positive alignment weight has been learnt, which makes them being aligned.

Figure 5 shows an alignment example of concept precision comparing the baseline with our proposed method. Although the ISI alignment incorrectly aligns the AMR concept token “thin(g)” to the English token “how,” our proposed method does not align them. Because “thin(g)” and “how” co-occurs several times in the training data, the ISI alignment gives it a high translation probability and aligns it. Our proposed

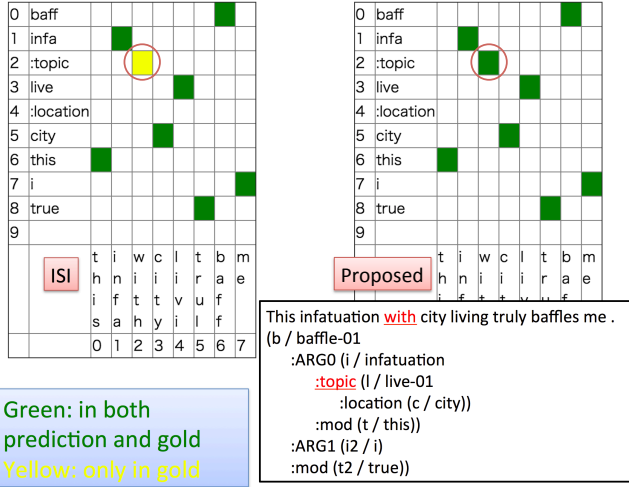


Figure 4: An alignment comparison example of better role recall.

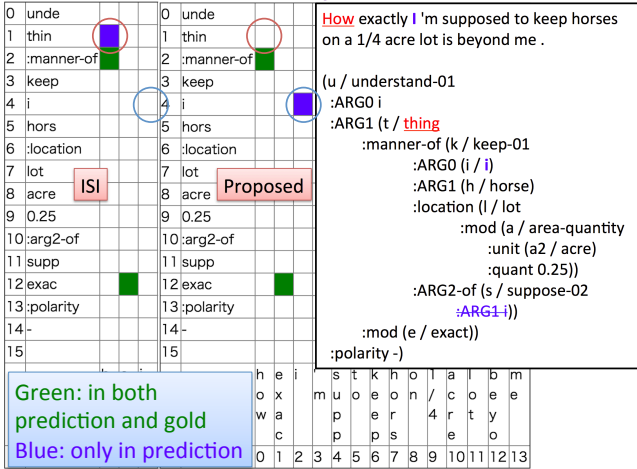


Figure 5: An alignment comparison example of concept pre-cision.

method also looks at the syntax information of “thin(g)” and “how,” which are “Sstatement” and “WHNP (WRB),” respectively. Because this syntax pair rarely co-occurs in the training data, the learnt weight is negative, which prevents this incorrect alignment. However, the proposed method incorrectly aligns the two “i” tokens in AMR and English. This happens because of two reasons: Firstly, the correct alignment “:ARG1 i” has been removed in the preprocessing; Secondly, Nile has a feature that tends to align same tokens.

### 5.3 AMR Parsing Results

The AMR parsing results are shown in Table 5, where “Our split” denotes the performance for different parsers on our data split of the AMR corpus.<sup>12</sup> For reference, we

<sup>12</sup>As the parsers of [Werling *et al.*, 2015; Zhou *et al.*, 2016; Peng *et al.*, 2017] are not publicly available, we could not report the performance of their parser on our data split. We were not able to run the parsers of [Artzi *et al.*, 2015; Misra and Artzi, 2016] on our split.

Method	Our split	Original split
[Pust <i>et al.</i> , 2015] (ISI)	64.7%	65.4%
[Pust <i>et al.</i> , 2015] (Proposed)	<b>65.1%</b>	N/A
[Pust <i>et al.</i> , 2015] (ISI) w/ rules	N/A	<b>67.1%</b>
[Flanigan <i>et al.</i> , 2014]	59.0%	58.2%
[Wang <i>et al.</i> , 2015]	61.3%	63.0%
[Zhou <i>et al.</i> , 2016]	N/A	66.0%

Table 5: Smatch F-scores for AMR parsing.

also list the parsing accuracies on the original split reported in [Pust *et al.*, 2015; Flanigan *et al.*, 2014; Wang *et al.*, 2015; Zhou *et al.*, 2016] in the “Original split” column of Table 5.<sup>13</sup> “[Pust *et al.*, 2015] (ISI)” and “[Pust *et al.*, 2015] (Proposed)” denote the parsers using different alignment models. “[Pust *et al.*, 2015] (ISI) w/ rules” denotes a system that further used rule-based alignments,” which is not a comparison object in our study. Note that we cannot report the performance of “[Pust *et al.*, 2015] (Proposed)” on the original split, because our proposed alignment method requires the 200 pairs in the development and testing sets of the original split for training and tuning the alignment model.

Parsing accuracies were evaluated using the Smatch F-score [Cai and Knight, 2013]. As reported [Cai and Knight, 2013], the Smatch F-score of human inter-annotator is in the 79-83 range. We can that see on both our and the original split, [Pust *et al.*, 2015] outperforms the other studies. On our split, the improved alignments by our proposed method also lead to a 0.4% Smatch F-score improvement for AMR parsing with the parser of [Pust *et al.*, 2015]. The performance of “[Pust *et al.*, 2015] (ISI)” differs on “Our split” and “Original split.” The reason for this is twofold: Firstly, “Original split” uses 100 more pairs for tuning the parser; Secondly, the 100 testing pairs moved from the testing data of the original split might be more difficult for parsing, which is consistent with the parsing performance of [Wang *et al.*, 2015].

## 6 Conclusion

The alignment between English sentences and AMR graphs is necessary for AMR parsing. We improved the alignment accuracy with a supervised syntax-based alignment method. We showed the effectiveness of the supervised method on both alignment and AMR parsing, even when only a very small training data set is available (i.e., 100 pairs).

As future work, firstly, we plan to increase the number of AMR/English sentence pairs with gold alignments for training a more accurate alignment model. Secondly, we plan to improve the alignment accuracy for roles. Semantic role labeling [Gildea and Jurafsky, 2000] for the English sentences

<sup>13</sup> Note that [Flanigan *et al.*, 2014] did not report the result on this dataset in their paper. [Pust *et al.*, 2015] reported the parsing performance of the parser of [Flanigan *et al.*, 2014] on this dataset, which we listed here. As [Werling *et al.*, 2015; Artzi *et al.*, 2015; Misra and Artzi, 2016] only reported the parsing performance (62.2%, 66.2%, and 66.0% Smatch F-score, respectively) on the newswire section of the original split, [Peng *et al.*, 2017] reported their performance (52% Smatch F-score) on a slightly larger dataset (LDC2015E86), we do not list their scores in Table 5.

and using the obtained roles for alignment may be a solution for this.

## Acknowledgments

We are very appreciated to Mr. Michael Pust for providing the AMR to constituency tree conversion code and helping conduct the AMR parsing experiments on his parser. We also thank Mr. Yevgeniy Puzikov very much for helping conduct the AMR parsing experiments on the JAMR and CAMR parsers.

## References

- [Artzi *et al.*, 2015] Yoav Artzi, Kenton Lee, and Luke Zettlemoyer. Broad-coverage ccg semantic parsing with amr. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1699–1710, Lisbon, Portugal, September 2015.
- [Banarescu *et al.*, 2013] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2013.
- [Brown *et al.*, 1993] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312, 1993.
- [Cai and Knight, 2013] Shu Cai and Kevin Knight. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria, August 2013.
- [Collins, 2002] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8, July 2002.
- [Flanigan *et al.*, 2014] Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland, June 2014.
- [Gildea and Jurafsky, 2000] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong, October 2000.
- [Misra and Artzi, 2016] Dipendra Kumar Misra and Yoav Artzi. Neural shift-reduce ccg semantic parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1775–1786, Austin, Texas, November 2016.
- [Och and Ney, 2003] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.
- [Palmer *et al.*, 2005] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, March 2005.
- [Peng *et al.*, 2017] Xiaochang Peng, Chuan Wang, Daniel Gildea, and Nianwen Xue. Addressing the data sparsity issue in neural amr parsing. In *Proceedings of the 15th European Chapter of the Association for Computational Linguistics*, Valencia, Spain, April 2017.
- [Petrov and Klein, 2007] Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April 2007.
- [Pourdamghani *et al.*, 2014] Nima Pourdamghani, Yang Gao, Ulf Hermjakob, and Kevin Knight. Aligning english strings with abstract meaning representation graphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 425–429, Doha, Qatar, October 2014.
- [Pust *et al.*, 2015] Michael Pust, Ulf Hermjakob, Kevin Knight, Daniel Marcu, and Jonathan May. Parsing english into abstract meaning representation using syntax-based machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1143–1154, Lisbon, Portugal, September 2015.
- [Riesa *et al.*, 2011] Jason Riesa, Ann Irvine, and Daniel Marcu. Feature-rich language-independent syntax-based alignment for statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 497–507, Edinburgh, Scotland, UK., July 2011.
- [Wang *et al.*, 2015] Chuan Wang, Nianwen Xue, and Sameer Pradhan. Boosting transition-based amr parsing with refined actions and auxiliary analyzers. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 857–862, Beijing, China, July 2015.
- [Werling *et al.*, 2015] Keenon Werling, Gabor Angeli, and Christopher D. Manning. Robust subgraph generation improves abstract meaning representation parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 982–991, Beijing, China, July 2015.
- [Zhang *et al.*, 2004] Ying Zhang, Stephan Vogel, and Alex Waibel. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal, May 2004. ACL Anthology Identifier: L04-1489.

[Zhou *et al.*, 2016] Junsheng Zhou, Feiyu Xu, Hans Uszkor-eit, Weiguang QU, Ran Li, and Yanhui Gu. Amr parsing with an incremental joint model. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 680–689, Austin, Texas, November 2016.